(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENTAMT

(12) **Offenlegungsschrift**

(10) **DE 43 04 082 A1**

(51) Int. Cl.[5]:
G 06 K 9/00
G 06 K 11/00

(21) Aktenzeichen: P 43 04 082.9
(22) Anmeldetag: 11. 2. 93
(43) Offenlegungstag: 18. 8. 94

(71) Anmelder:
Nitzschmann, Bernd, Dipl.-Phys., 26121 Oldenburg, DE

(72) Erfinder:
gleich Anmelder

(56) Für die Beurteilung der Patentfähigkeit in Betracht zu ziehende Druckschriften:

DE 41 19 091 A1
SU 15 43 431 A1
SCHÜRMANN,Jürgen: Automatisches Lesen. In: net nachrichten elektronik + telematik 36, 1982,H.11, S.473-477;
RICHTER,Werner: Datenerfassung -Tastatur ade? In: Der Elektroniker 3/88,S.32-34;
JP 63-206882 A. In: Patents Abstracts of Japan, P-806,Dec.26,1988,Vol.12,No.497;

(54) Verfahren zum verbesserten Erkennen von gedruckten Schriftzeichen

(57) Durch Anwendung von drei verschiedenartig arbeitenden Computer-Programmen oder Algorithmen zur optischen Zeichenerkennung (OCR) auf eine Textvorlage mit anschließendem zeichenweisen Vergleich der Resultate, wobei im Falle von Unterschieden zwischen den drei Resultaten das in den Ausgabetext zu übernehmende Zeichen durch 2 : 1-Abstimmung gefunden wird, ergibt sich eine insgesamt höhere Wiedererkennungsrate von gedruckten Texten durch Computer-Programme und damit ein geringerer Aufwand bei eventueller manueller Nachkorrektur.

DE 43 04 082 A1

DE 43 04 082 A1

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen

BUNDESDRUCKEREI

Beschreibung

Stand der Technik

Es sind verschiedene Verfahren zum Erkennen von Texten aus Bildvorlagen mit Hilfe von EDV-Anlagen bekannt. Ihnen ist gemeinsam, daß auf die aus Punktmustern bestehenden Buchstaben der Bildvorlage ein Algorithmus angewendet wird, um aus diesen Punktmustern den zugeordneten Buchstaben mit möglichst großer Sicherheit zu bestimmen.

Das Grundproblem bei der Bestimmung der in Druckvorlage enthaltenen Zeichen besteht darin, daß grundsätzlich eine Ähnlichkeitsbestimmung der Punktmuster nach vorzugebenden Kriterien durchgeführt werden muß, da auch bei qualitativ hochwertigen Vorlagen alle in der Vorlage vorhandenen Buchstaben verschiedene Punktmuster in der Bilddatei ergeben.

Die Aufgabe von OCR-Verfahren ist es gleichzeitig, die Erkennung einer Vielfalt von Schriftarten (möglichst simultan) zu ermöglichen, so daß bei einem gegebenen Punktmuster kein eindeutiger Bezug auf vorgegebene Muster möglich ist.

Dieser Umstand verhindert eine exakte Erkennung des Originaltextes mit 100%iger Erkennungssicherheit und damit auch ein eindeutiges Vorgehen bei der optischen Texterkennung.

Aus dieser Tatsache heraus sind viele verschiedene Verfahren zur optischen Texterkennung mittels Software entwickelt worden, die alle unterschiedliche Charakteristika bei der Erkennung aufweisen.

Der Ähnlichkeitsgrad, den der jeweilige Algorithmus ausgibt, ist damit auch von diesem abhängig und gibt nur eine relative Ähnlichkeit bezogen auf die jeweils verwendeten Vergleichskriterien an.

Unter den existierenden Verfahren befinden sich hybride Verfahren, die versuchen, ein Punktmuster mit Hilfe eines Algorithmus zu identifizieren und die im Falle einer zu geringen Wiedererkennungssicherheit (die ja nur relativ angegeben werden kann) weitere Erkennungsalgorithmen zur Entscheidungsfindung heranzuziehen.

Aufgabe

Aufgabe der Erfindung ist es, ein Verfahren vorzuschlagen, das eine Reproduzierung von gedruckten Texten in EDV-Anlagen mittels Software mit größerer Genauigkeit (gemeint ist die prozentuale Übereinstimmung des reproduzierten Textes bezogen auf die gedruckte Textvorlage) als bisher üblich ermöglicht.

Verfahren

Die Grundidee des Verfahrens beruht auf der gleichzeitigen Anwendung von drei möglichst verschiedenartig gestalteten Erkennungsprozessen auf eine Textvorlage. Mit verschiedenartig gestaltet ist gemeint, daß sich die drei Erkennungsprozesse durch den Erkennungsalgorithmus und/oder die programmtechnische Ausführung des Algorithmus und/oder die für den Algorithmus notwendigen Hilfsparameter unterscheiden müssen.

Bei einem anschließenden synchronisierten Vergleich der gelieferten Ausgaben wird aufgrund des Vorhandenseins von drei OCR-Resultaten an jenen Stellen, an denen eine OCR-Ausgabe von den beiden anderen OCR-Ausgaben verschieden ist, angenommen, daß die

zwei gleichen Textstellen dem Originaltext entsprechen. Die beim dritten Prozeß entstandene zu den beiden anderen Textstellen verschiedene Textstelle wird verworfen.

Ein derartiges Vorgehen ist nur bei mindestens drei und einer ungeraden Anzahl von OCR-Resultaten möglich. Es wird dabei implizit eine gleich große Erkennungssicherheit aller drei Erkennungs-Methoden angenommen.

Der Vorteil des vorgeschlagenen Verfahrens besteht darin, daß auf jedes Punktmuster drei verschiedene Bewertungskriterien angewendet werden. Falls dann ein Kriterium aufgrund der problembedingten relativen Genauigkeit einen falschen Buchstaben vorhergesagt, zeigt die Erfahrung, daß in den meisten solcher Fälle die beiden anderen Kriterien die "richtige" Vorhersage treffen, bei der anschließenden Synthese wird dann auch bei der 2 : 1 Abstimmung das "richtige" Zeichen geliefert. Zusätzlich kann an Textstellen, an denen alle drei Verfahren unterschiedliche Angabe machen, mit höherer Sicherheit als bei Verwendung eines Verfahrens davon ausgegangen werden, daß in der Original-Vorlage die entsprechende Stelle für OCR-Automaten nicht erkennbar war, z. B. durch Verschmutzung, Ungenauigkeiten im Druck, etc.

Im Gegensatz dazu stehen die o. g. hybriden Verfahren, die aus Gründen der Rechenzeitersparnis ein bestimmtes Verfahren primär einsetzen und weitere Verfahren zur Erkennung eines bestimmten Musters nur dann, falls der primäre Algorithmus eine geringe Wiedererkennungssicherheit angibt.

Differieren alle drei OCR-Resultate an einer bestimmten Stelle, wird ein Vorgehen entsprechend der Unteransprüche 2 ... 5 vorgeschlagen.

Um drei OCR-Ausgabetexte synchron vergleichen zu können, wird ein als Computerprogramm realisierter Algorithmus verwendet, der jeden OCR-Text mit den beiden jeweils anderen Texten zeichenweise vergleicht und an den Positionen, an denen die beiden jeweils verglichenen Texte einen Unterschied aufweisen, gleichzeitig die Differenztexte bestimmt und eine Resynchronisation erreicht, z. B. falls der Textunterschied in zusätzlichen (oder fehlenden) Buchstaben besteht.

Dabei wird ein wiederholtes versuchsweise gleichzeitiges Entfernen von Zeichenketten variierender Länge aus den beiden zu vergleichenden Texten ab der Position des ersten verschiedenen Zeichens und Speichern a) der Zeichenketten und b) der dadurch erzielten Übereinstimmung der Texte ab der Verschiebungsposition vorgenommen, bei anschließender Auswahl derjenigen zwei versuchsweise entfernten Zeichenketten, die eine maximale Übereinstimmung der restlichen Texte zur Folge haben.

Patentansprüche

1. Verfahren zum verbesserten Erkennen von gedruckten Schriftzeichen mit Hilfe von in EDV-Anlagen ablaufenden Computer-Programmen und Übertragung der Schriftzeichen in eine in EDV-Anlagen übliche Repräsentation von Texten (z. B. ADCII), **gekennzeichnet durch** die gleichzeitige Verwendung von drei verschiedenen Programmen oder Algorithmen zur optischen Zeichenerkennung (optical character recognition OCR) und synchronem Zusammenführen der drei dadurch erhaltenen Texte zu einem Text, dergestalt, daß an solchen Textstellen, an denen sich die drei OCR-Vorlagen

unterscheiden, ein als Computerprogramm realisiertes Verfahren angewendet wird, das aufgrund des Vorhandenseins von drei Textvorlagen den in den Ausgabetext zu übernehmenden Textteil bestimmt.

2. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß mehr als 3 verschiedene OCR-Verfahren auf eine Druckvorlage angewendet werden und das die entsprechende Anzahl von OCR-Ausgabetexten zu einem Resultat-Text zusammengeführt wird.

3. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß beim Zusammenführen der Texte an denjenigen Positionen, an denen alle drei Texte voneinander abweichen, zusätzliche OCR-Verfahren zur Bestimmung des Ausgabetextes aufgerufen werden.

4. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß beim Zusammenführen der Texte an solchen Positionen, an denen alle drei Texte voneinander abweichen, die Möglichkeit der manuellen Texteingabe und sonstiger Einflußnahme auf das weitere Programmverhalten gegeben ist.

5. Verfahren nach Anspruch 4, dadurch gekennzeichnet, daß nach manueller Texteingabe eine erneute Synchronisation der drei durch OCR-Verfahren erhaltenen Texte durchgeführt wird.

6. Verfahren nach Anspruch 4, dadurch gekennzeichnet, daß eine evtl. manuelle Nachkorrektur erst nach vollständigem Zusammenführen der drei OCR-Resultate bei vorläufiger Auslassung der für manuelle Korrektur vorgemerkten Textstellen durchgeführt wird.

– Leerseite –